Statistical Data Mining and Machine Learning (full question)

6. You are given a dataset $\{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. Assume that the sample covariance matrix $S = \frac{1}{n-1}\sum_{i=1} x_i x_i^\top$ has eigendecomposition $S = V\Lambda V^\top$, where $\Lambda = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a diagonal matrix containing eigenvalues of S in decreasing order.

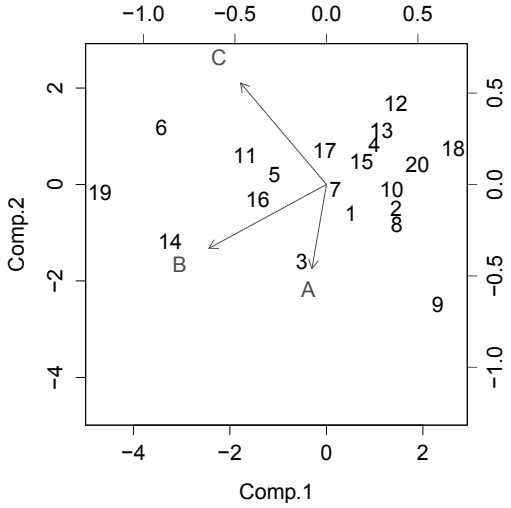
   (a) Using the notation above, define the projection of x_i onto the first k principal components. What is the proportion of variance explained by these first k principal components? Carefully define any additional notation you introduce.

   (b) Assume that you have $n = 20$ observations with $p = 3$, and that the original variables are called `A`, `B`, and `C`.

   The biplot on the right is *unscaled*, i.e. it is obtained using R commands

   ```
   x.pca <- princomp(x)
   biplot(x.pca, scale=0)
   ```

   How is the position of the letter `A` in this biplot determined from the matrix V? Read off the approximate first eigenvector of S from the biplot.
   [*Hint: Recall that the top and the right axes in the biplot correspond to the projections of the original variables*]



   (c) You are given a probabilistic PCA (PPCA) model:

$$\begin{aligned} Y_i &\sim \mathcal{N}(0, I_k), \\ X_i|Y_i &\sim \mathcal{N}\left(LY_i, \sigma^2 I_p\right), \end{aligned}$$

   where $\{Y_i\}_{i=1}^n$ are k-dimensional latent random vectors, $\theta = (L, \sigma^2)$ are the parameters with $L \in \mathbb{R}^{p\times k}$ and $\sigma^2 > 0$, and $\{X_i\}_{i=1}^n$ are p-dimensional observed random vectors.

   (i) What is the resulting model for the marginal distribution of X_i?

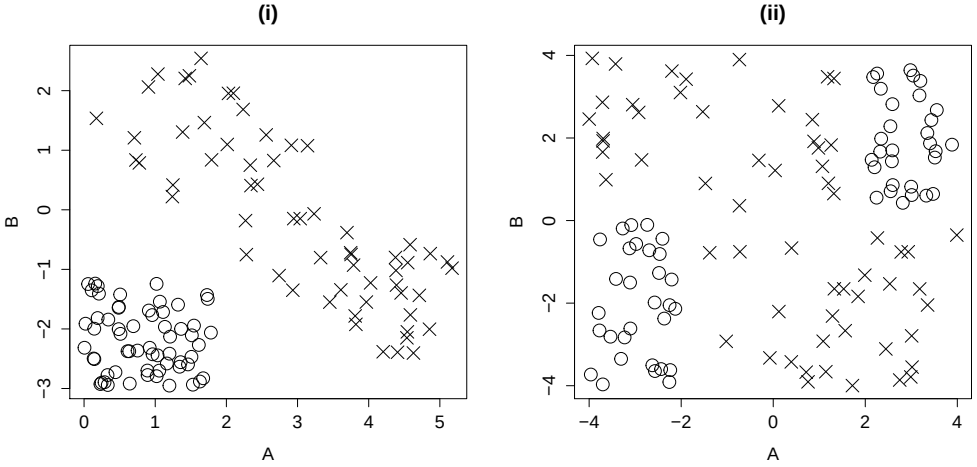
   (ii) Both PCA and PPCA aim to recover the latent k-dimensional representation of data. Briefly comment on how the recovered representations differ between PCA and PPCA.

   (iii) Assume that you are given the E-step of EM algorithm for PPCA with the corresponding variational distribution over all latent vectors $\mathbf{y} = [y_1, \ldots, y_n]$ being a product of independent normal distributions, $q(\mathbf{y}) = \prod_{i=1}^n \mathcal{N}(y_i; b_i, R)$, for some mean vectors $b_i \in \mathbb{R}^k$ and a common covariance matrix $R \in \mathbb{R}^{k\times k}$. Derive the M-step for L as a function of R, $\{b_i\}_{i=1}^n$ and the observations $\{x_i\}_{i=1}^n$.
   [*Hint: You can use matrix derivatives* $\frac{\partial a^\top W b}{\partial W} = ab^\top$, $\frac{\partial \text{Tr}[W^\top W A]}{\partial W} = WA^\top + WA$.]

Statistical Data Mining and Machine Learning (full question)

7. (a) Which of the following classifiers are generative: linear discriminant analysis, logistic regression, support vector machine, naïve Bayes? Name one advantage of generative classifiers over discriminative ones.

   (b) You are given points from two classes: "○" and "×", which are described in terms of variables `A` and `B`. For each of the sets of points in the figures below, draw a decision tree of depth 2 that can separate the given data completely (by using thresholding of *a single variable* in each node).



   (c) Consider a binary classification problem with classes denoted -1 and $+1$. For a *soft classification rule* $f : \mathbb{R}^p \to \mathbb{R}$, the exponential loss is defined as $L(y, f(x)) = e^{-yf(x)}$.

      (i) Briefly comment on the difference between the exponential and the logistic loss in terms of how they treat the misclassified examples.

      (ii) Show that the optimal classification rule, i.e. f^* which minimises the risk $R(f) = \mathbb{E}_{XY} e^{-Yf(X)}$, is given by

$$f^*(x) = \frac{1}{2} \log \frac{\mathbb{P}(Y = +1 | X = x)}{\mathbb{P}(Y = -1 | X = x)}.$$

      (iii) Consider the linear decision boundary $f(x) = w^\top x$. Write down the objective function $J(w)$ given by the L_2-regularised empirical risk with respect to the exponential loss. Derive the gradient descent update rule for w.